Resource for the exploration of regulons accurately predicted by the methods of comparative genomics.

Pavel S. Novichkov[1], Mikhail S. Gelfand[2], Inna Dubchak[1,4], Dmitry A. Rodionov[2,3]

[1]*Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA;* [2]*Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow 127994, Russia;* [3]*Burnham Institute for Medical Research, La Jolla, CA 92037,USA;* [4]*Department of Energy Joint Genome Institute, Walnut Creek, CA 94598, USA.*

Identification and reconstruction of various transcriptional regulons in bacteria using a computational comparative genomics approach is coming of age. During the past decade a large number of manually-curated high quality inferences of transcriptional regulatory interactions were accumulated for diverse taxonomic groups of bacteria. These data provide a good foundation for understanding molecular mechanisms of transcriptional regulation, identification of regulatory circuits, and interconnections among circuits within the cell. Traditional experimental methods for regulon analysis have certain limitations both in terms of productivity and feasibility. While the development of high-throughput transcriptome approaches allow to obtain genome-scale gene expression patterns, in many cases the complexity of the interactions between regulons makes it difficult to distinguish between direct and indirect effects on transcription. The availability of a large number of closely related genomes allows one to apply comparative genomics to accurately expand already known regulons to yet uncharacterized organisms, and to predict and describe new regulons. Due to fast accumulation of such valuable data, there is a need for a specialized database and associated analysis tools that will compile and present the growing collection of high quality predicted bacterial regulons.

The RegPrecise database was developed for capturing, visualization and analysis of transcription factor regulons that were reconstructed by the comparative genomic approach. The primary object of the database is a single regulon in a particular genome, which is described by the identified transcription factor, its DNA binding site model (a profile), as well as the set of regulated genes, operons and associated operator sites. Regulons for orthologous transcription factors from closely related genomes are combined into the collections that provide an overview of the conserved and variable components of the regulon. A higher level representation of the regulatory interactions is also provided for orthologous regulons described in several bacterial taxonomic groups enabling comparison and evolutionary analysis of the transcription factor binding motifs. Another view of complex data in the database is a general overview of multiple regulons inferred in a set of closely related group of genomes.

The current version of database covers more than 250 genomes and 180 profiles. Among others, it represents the results of our recent comparative genomic reconstruction of metabolic regulons in 13 Shewanella species that included near 70 transcription factors, approximately 400 binding sites and more than 1000 target genes per each genome. The database gives access to large regulatory networks reconstructed for certain metabolic pathways, e.g. degradation of fatty acids, branch chain amino acids, and aminosugars,

homeostasis of biometals, and biosynthesis of NAD cofactor. In the near future we are planning to add a large collection of regulons for the LacI family transcription factors.